University of Waterloo

A Timing-Driven Algorithm for Leakage Reduction in MTCMOS FPGAs

Hassan Hassan, Mohab Anis, and Mohamed Elmasry
University of Waterloo
VLSI Research Group

Outline

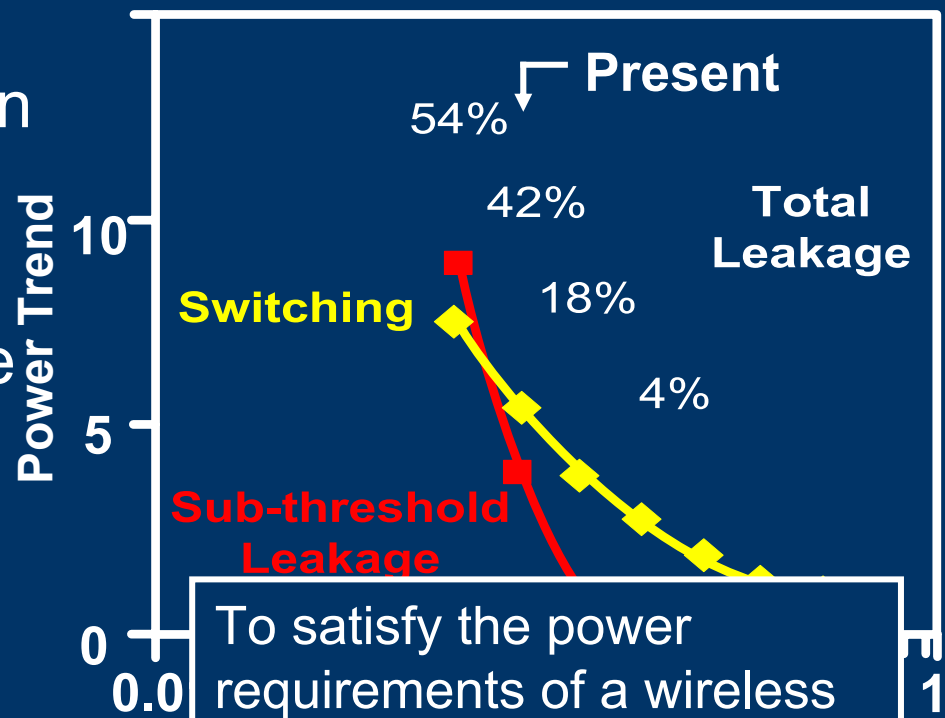
- Introduction and Motivation,
- Proposed FPGA Architecture/CAD Flow,
- Discharge Current Processing,
- LAP: Logic Activity Profiles,
- Timing-driven MTCMOS Packing,
- Results and Discussions,
- Conclusion.

Outline

- Introduction and Motivation,
- Proposed FPGA Architecture/CAD Flow,
- Discharge Current Processing,
- LAP: Logic Activity Profiles,
- Timing-driven MTCMOS Packing,
- Results and Discussions,
- Conclusion.

Leakage Power in FPGAs

- As the CMOS process shrinks, V_{DD} is scaled down to reduce dynamic power,
- V_{th} is also scaled down to improve the CMOS device switching speed,
- But with low V_{th}, sub-threshold leakage power [illegible]xponentially.



4X increase in leakage from 130nm

[illegible] (avg input data)	[illegible]st-case input data	Worst-case input data
18.9µW/CLB	-31.1%	+26.8%

To satisfy the power requirements of a wireless mobile application: only 20 CLBs can be used

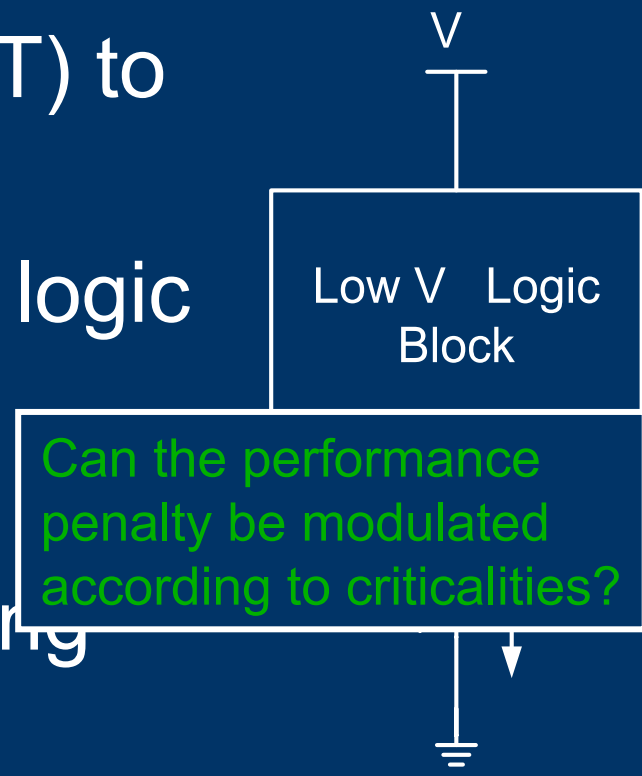
Increase in input dependency

The Status in Modern FPGAs

- Implemented in 65nm CMOS process.
- Average utilization per configuration in 60-70%.
- The unutilized parts represent a leakage power overhead without producing useful output:
  - For a utilization of 50%, 56% of the leakage power is consumed in the unutilized parts.
- Even the utilized parts consume active leakage in their active mode, and large standby leakage power during their idle period.

Supply Gating Architecture

- A high V_{th} sleep transistor (ST) to cut off the leakage path,
- The ST will turn OFF the idle logic blocks → static leakage,
- The ST reduces leakage significantly due to the stacking effect → dynamic leakage,
- Leakage reduction is traded to performance degradation.



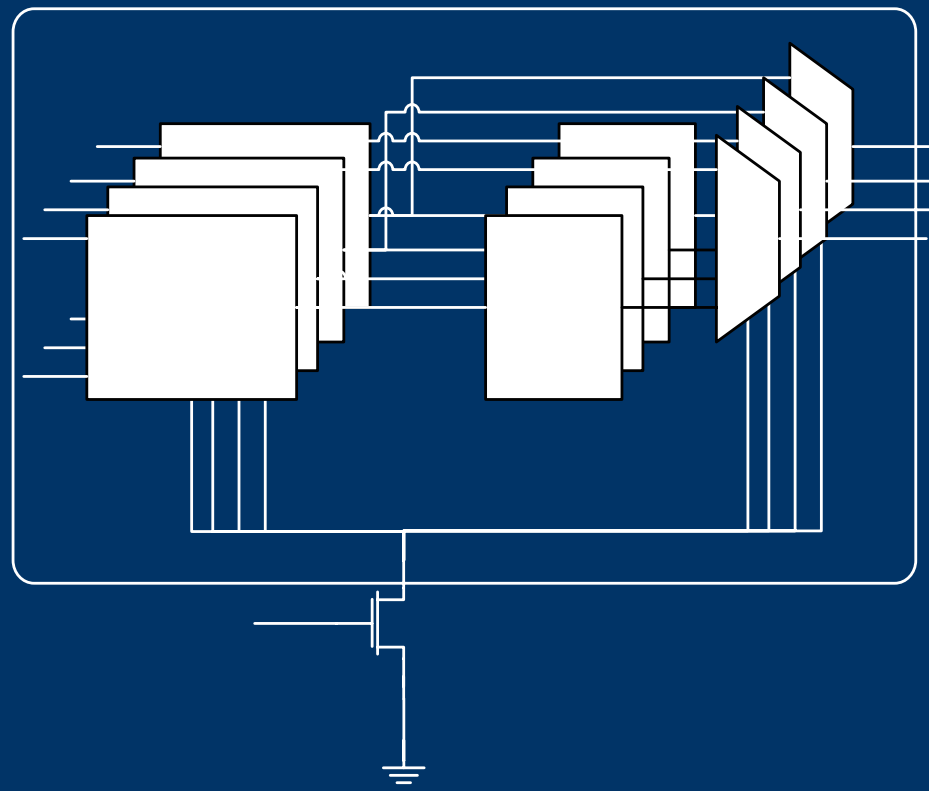
Performance penalty around 5% all over the chip.

Outline

- Introduction and Motivation,
- Proposed FPGA Architecture/CAD Flow,
- Discharge Current Processing,
- LAP: Logic Activity Profiles,
- Timing-driven MTCMOS Packing,
- Results and Discussions,
- Conclusion.

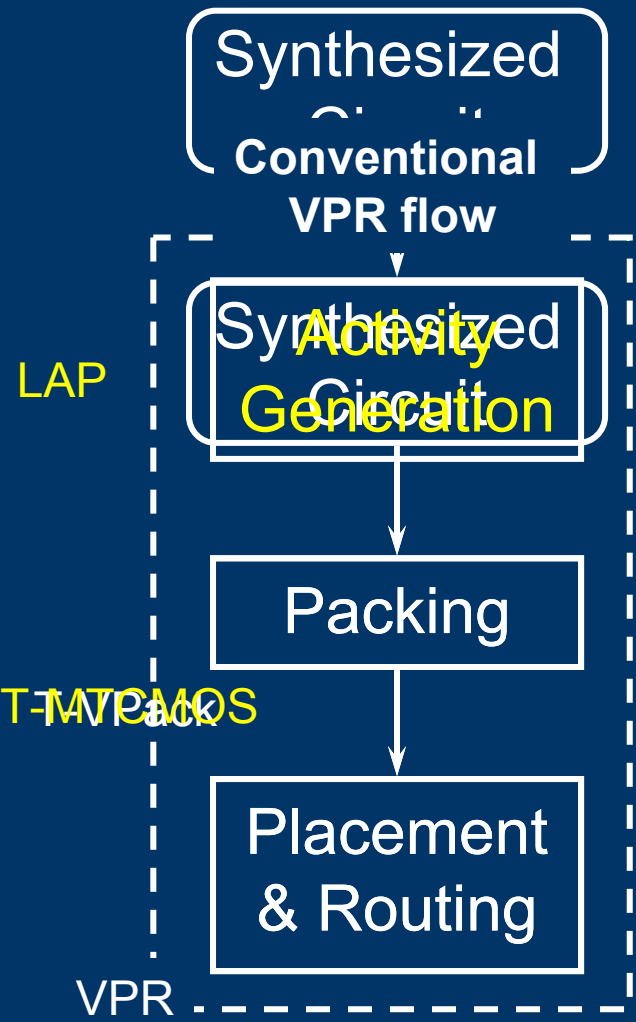
Targeted FPGA Architecture

- Every *n* BLEs are grouped together in one sleep region,
- Every sleep region is served by one ST,
- DFF are not supply gated and used for data retention during the sleep mode.



CAD Flow

- Conventional CAD flows for FPGAs do not target leakage power reduction,
- Identifying logic blocks that can be turned OFF simultaneously, activity profile generation,
- T-MTCMOS; a timing-activity modification of the T-VPack algorithm.



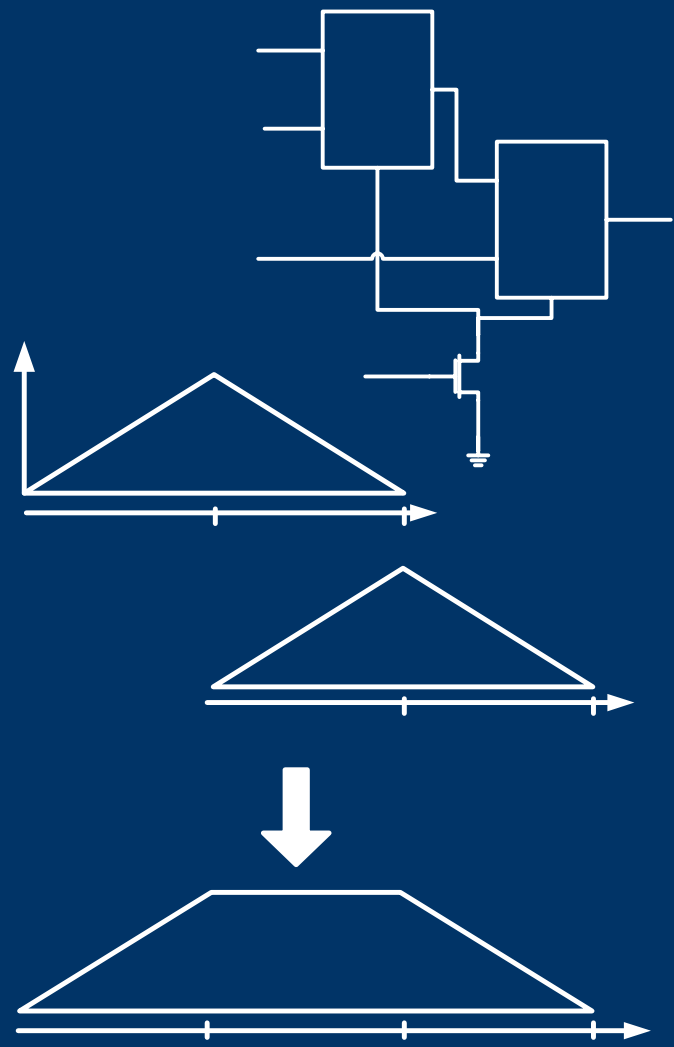
Outline

- Introduction and Motivation,
- Proposed FPGA Architecture/CAD Flow,
- Discharge Current Processing,
- LAP: Logic Activity Profiles,
- Timing-driven MTCMOS Packing,
- Results and Discussions,
- Conclusion.

Sizing the Sleep Transistor

- Performance Degradation:
  - The maximum discharge current flowing through the ST is limited by its size,
  - The total discharge current in any sleep region must be less than that of the ST,
  - Usually the performance penalty is limited to 5%,
  - Discharge current patterns depend on the connections of logic blocks.

$$\left.\frac{W}{L}\right|_{\text{sleep}} = \frac{I_{\text{sleep}}}{x\mu_n C_{ox}(V_{DD}-V_{thL})(V_{DD}-V_{thH})}$$



Outline

- Introduction and Motivation,
- Proposed FPGA Architecture/CAD Flow,
- Discharge Current Processing,
- LAP: Logic Activity Profiles,
- Timing-driven MTCMOS Packing,
- Results and Discussions,
- Conclusion.

LAP: Logic Activity Profiles

- The activity of each BLE is represented as a binary sequence (activity vector),
- The relation between the activities is calculated based on the Hamming distance between activity vectors,
- Activity vectors are generated based on the logic function of the blocks.

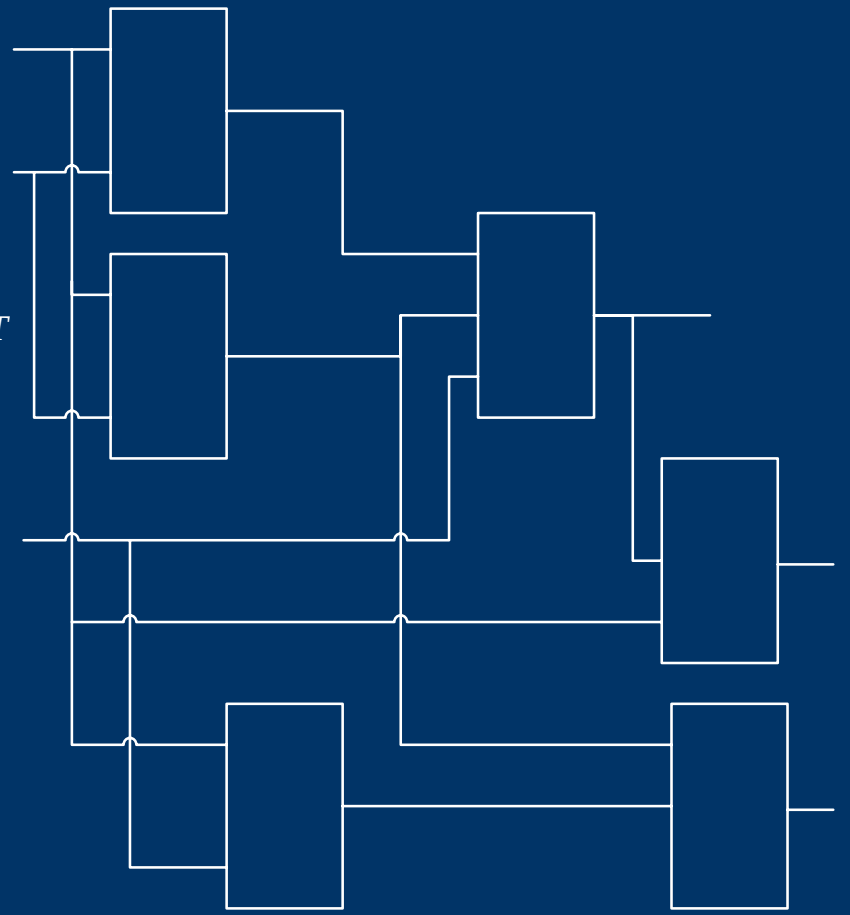
LAP: Logic Activity Profiles

- Activity vector
  - Given a net *x* in a circuit netlist, the activity vector A_x of *x* is

$$A_x = \begin{pmatrix} a_1 & a_2 & a_3 & .... & a_{2^{n-1}} & a_{2^n} \end{pmatrix}^T$$

  - *n* is the total number of inputs to the circuit
  - a_i is a binary number and equal to '1' if at input vector *I*, *x* is needed to evaluate any of the outputs
  - For example:

$$A_D = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}^T$$



LAP: Logic Activity Profiles

- Hamming distance
  - The Hamming distance between any 2 binary vectors:

$$d_{(a,b)} = \sum_{k=0}^{n-1} |a_k - b_k|$$

  - The difference between the activity vectors can be represented as the Hamming distance between them,
  - Hence, if *D* and *I* are grouped in the same cluster, the ST will be off for 25% of the time.

$$A_D = (0 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1)^T$$

$$A_I = (1 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 1)^T$$

$d_{(d,i)} = 4$

Outline

- Introduction and Motivation,
- Proposed FPGA Architecture/CAD Flow,
- Discharge Current Processing,
- LAP: Logic Activity Profiles,
- Timing-driven MTCMOS Packing,
- Results and Discussions,
- Conclusion.

AT-VPack: Activity Packing

- Based on the T-VPack algorithm,
- For each cluster, a seed BLE is selected with the highest criticality,
- In T-VPack, BLEs are added based on:
  - Cluster size does not exceed cluster capacity,
  - Number of inputs does not exceed cluster inputs.
- Added constraint:
  - Cluster discharge current does not exceed maximum discharge current.
- Objective function of adding block B to cluster C:

$$(1-\alpha)\left[\lambda Criticality(B)+(1-\lambda)SharingGain(B,C)\right]+\alpha\frac{2^n-d(B,C)}{2^n}$$

T-MTCMOS: Timing Driven Packing

- The discharge current constraint is relaxed for non-critical paths,

$$\hat{I}_{sleep} = I_{sleep}\left[1+\delta\left(1-\frac{criticality(C)}{Max\_criticality}\right)\right]$$

- This will result in packing more blocks with closer activity profiles, thus more leakage savings,
- It should be noted that no new critical paths get created.

Outline

- Introduction and Motivation,
- Proposed FPGA Architecture/CAD Flow,
- Discharge Current Processing,
- LAP: Logic Activity Profiles,
- Timing-driven MTCMOS Packing,
- Results and Discussions,
- Conclusion.

Results and Discussions

Circuit	% of Unused Clusters	% Leakage Savings w/o T-MTCMOS	% Leakage Savings T-MTCMOS
alu4	4.5	22.9	50.13
apex2	2.48	20.7	46.87
apex4	2.16	19.1	41.96
		20.2	42.87
		18.9	40.02
		16.8	39.67
		22	51.34
dsip	4.7	21.2	44.05
elliptic	5.9	21.6	
ex1010	0.26	18.7	
ex5p	6.92		
frisc	1.56		
misex3	2.21		
pdc	0.78		
s298	8.32	28.3	64.57
s38417	5.6	25.4	34.39
s38584.1	1.56	18.9	39.96
seq	0	14.2	23.64
spla	3.6	18.3	39.43

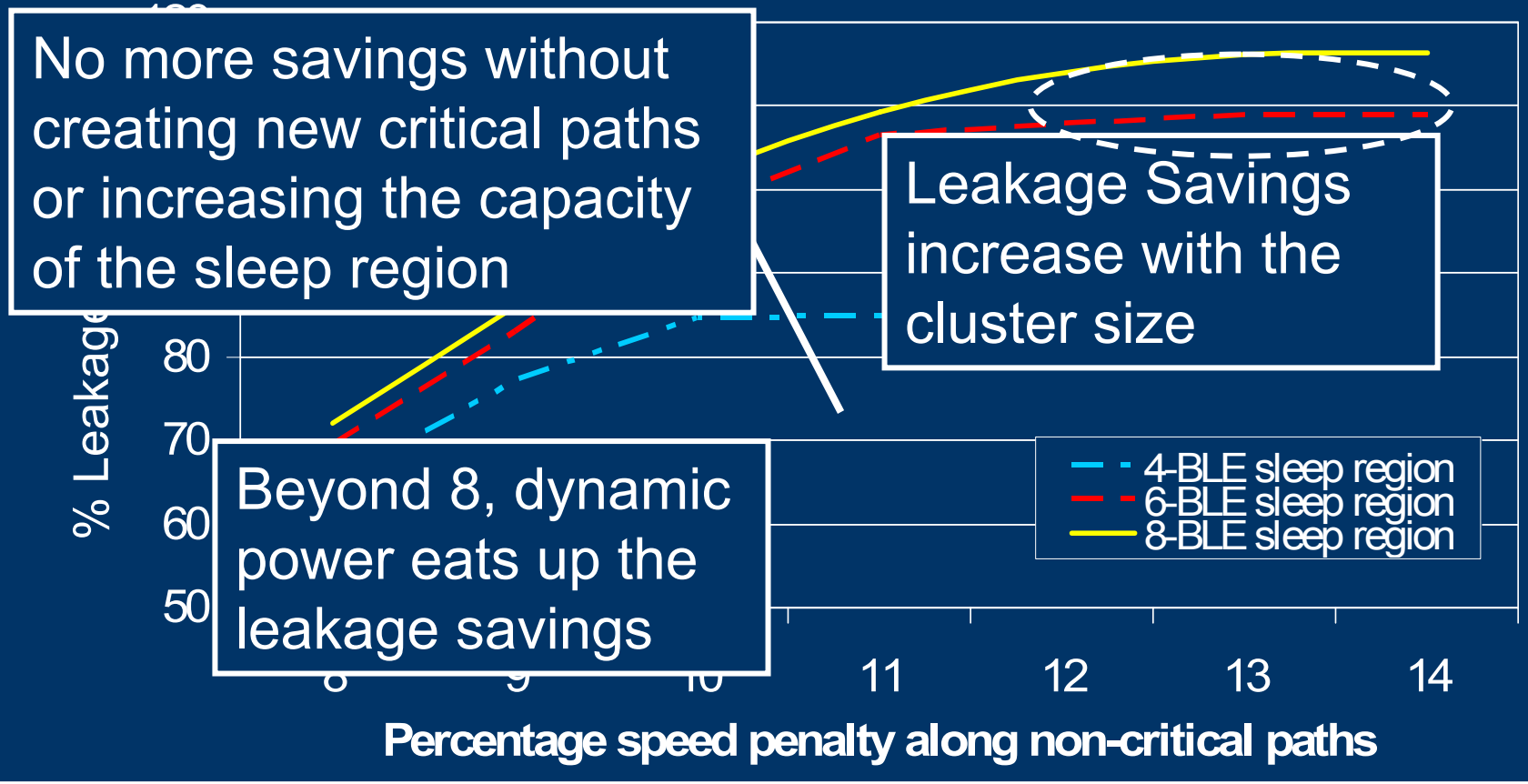
Almost 2X increase in leakage savings

Max unused CLBs
Max power savings

No unused CLE
Savings from the dynamic switching ON and OFF of the used CLBs

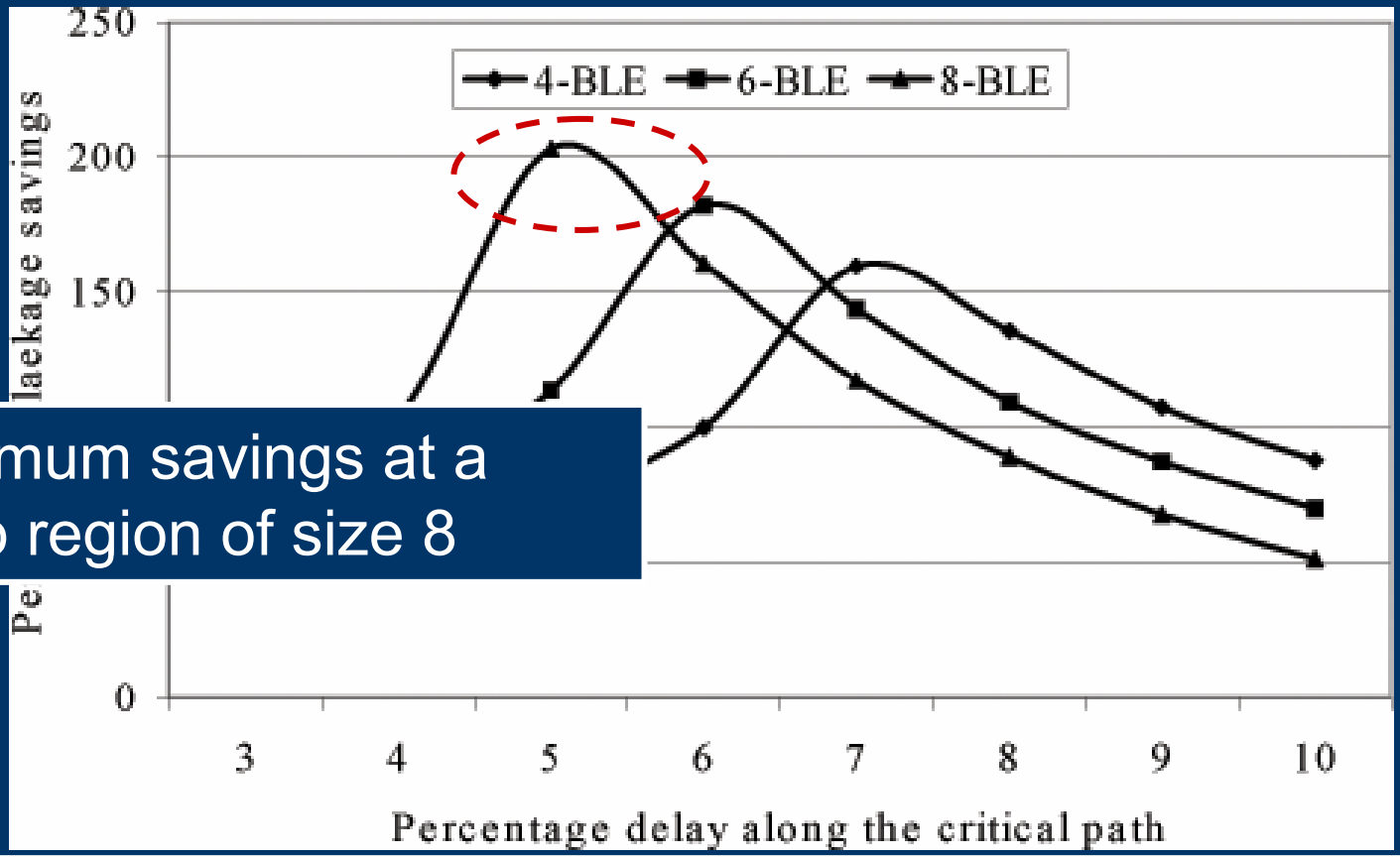
Results and Discussions

- Leakage Savings vs. non-critical paths sleep penalty:



Results and Discussions

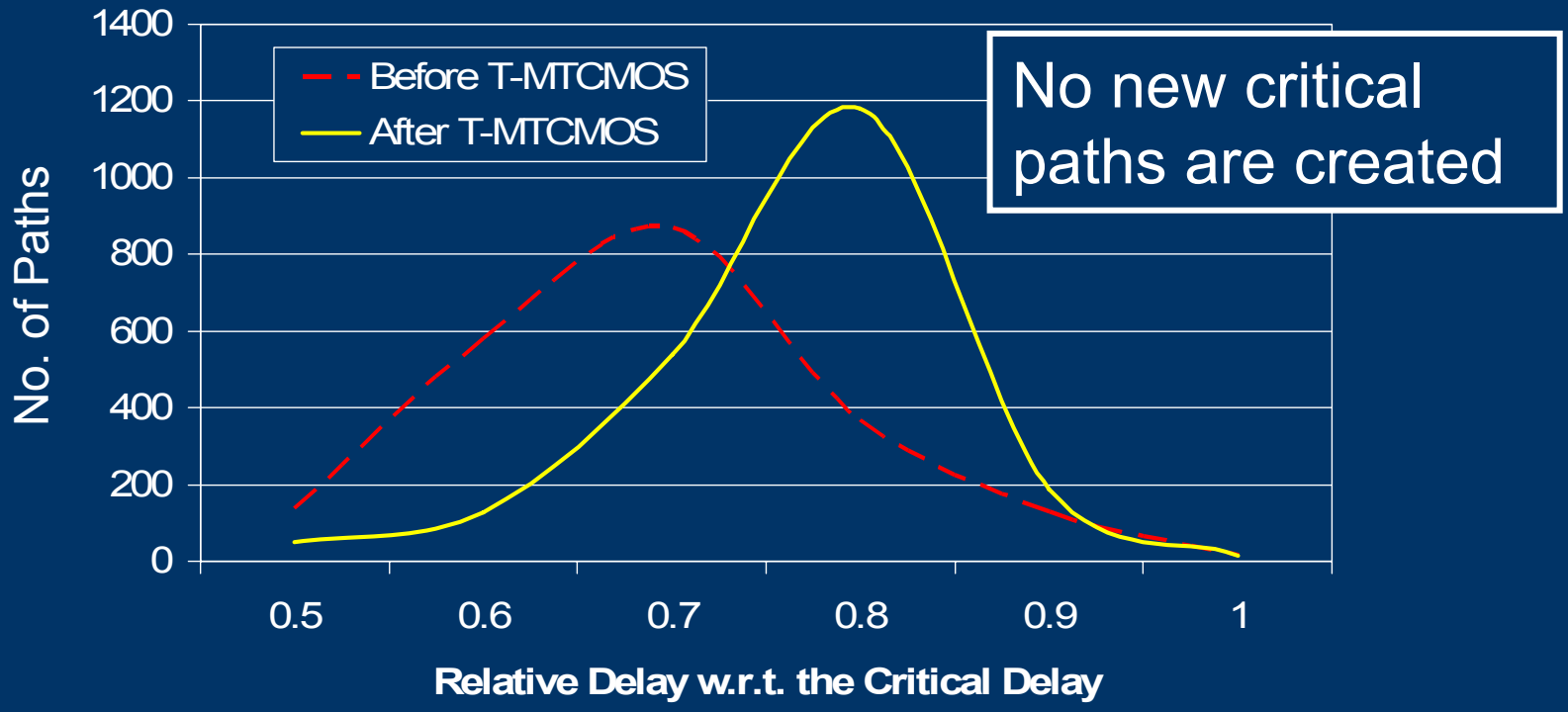
- Leakage Savings vs. critical paths sleep penalty:



Maximum savings at a sleep region of size 8

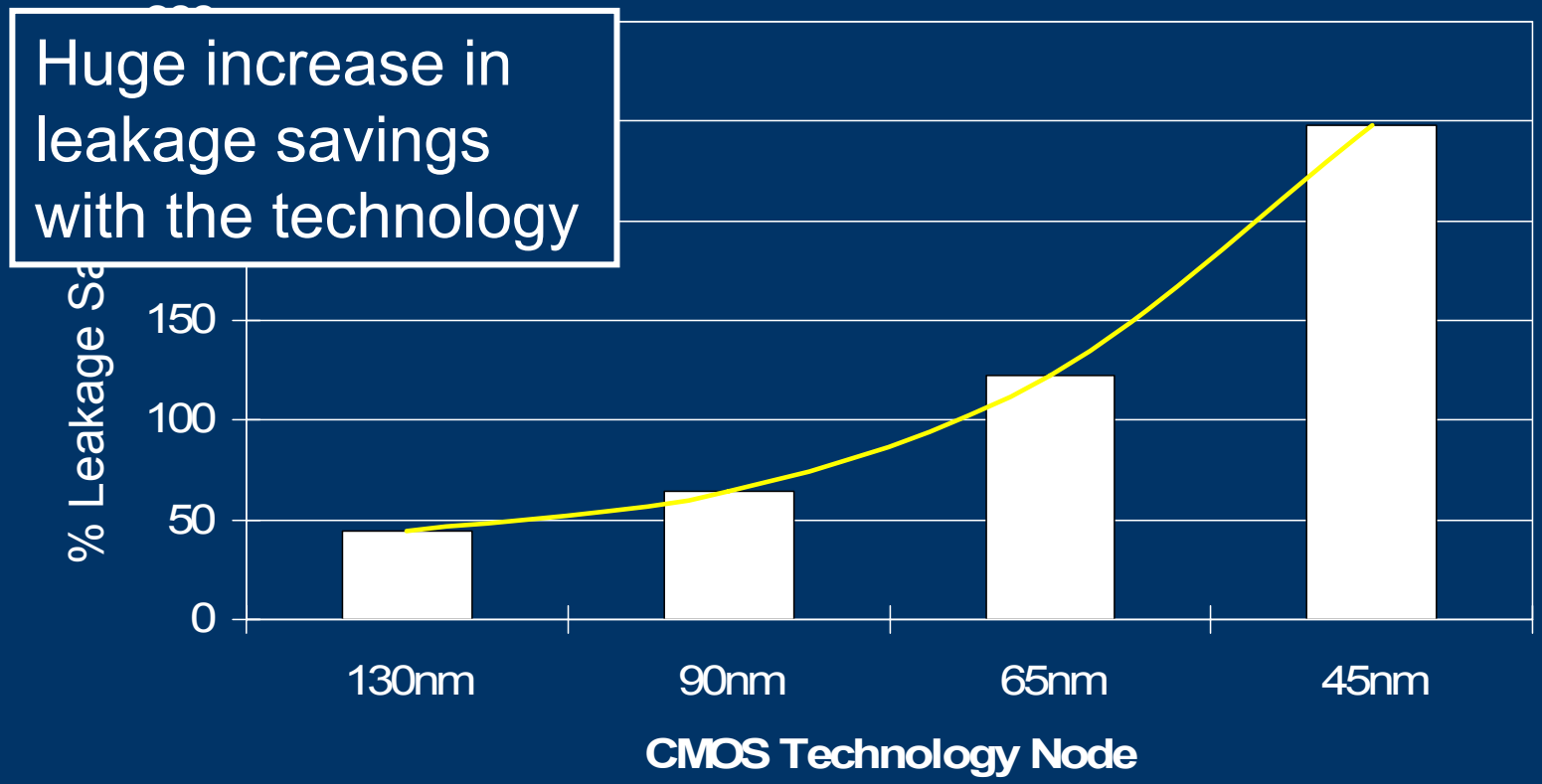
Results and Discussion

- Paths delay distribution:



Results and Discussion

- Leakage savings and technology scaling:



Conclusions

- Modulating the speed penalty due to sleep transistors in FPGAs results in a 2X increase in leakage savings,
- Leakage savings saturate with increasing the delay penalty along non-critical paths,
- A sleep region of size 8 results in the optimum leakage savings.